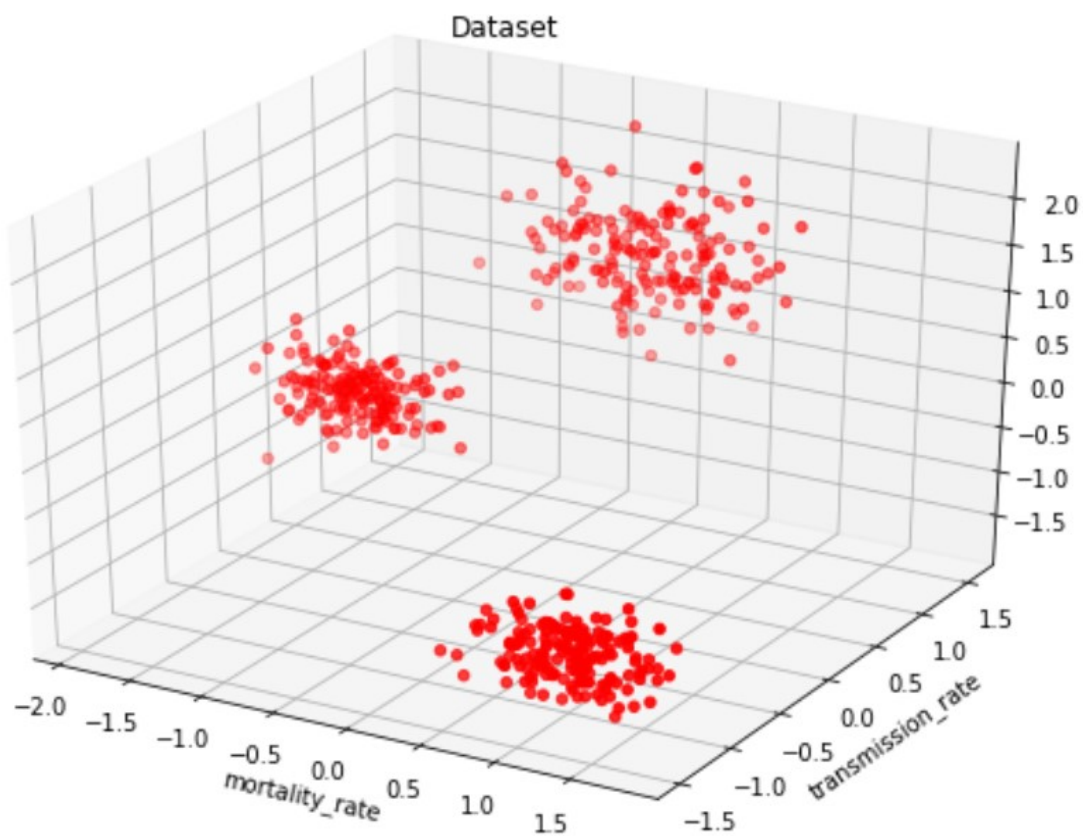


Vishal Sharma (19MF3IM26)

# Coronavirus Data Clustering using Complete Linkage Hierarchical Clustering Technique

---



*Scatter plot of the given Dataset*

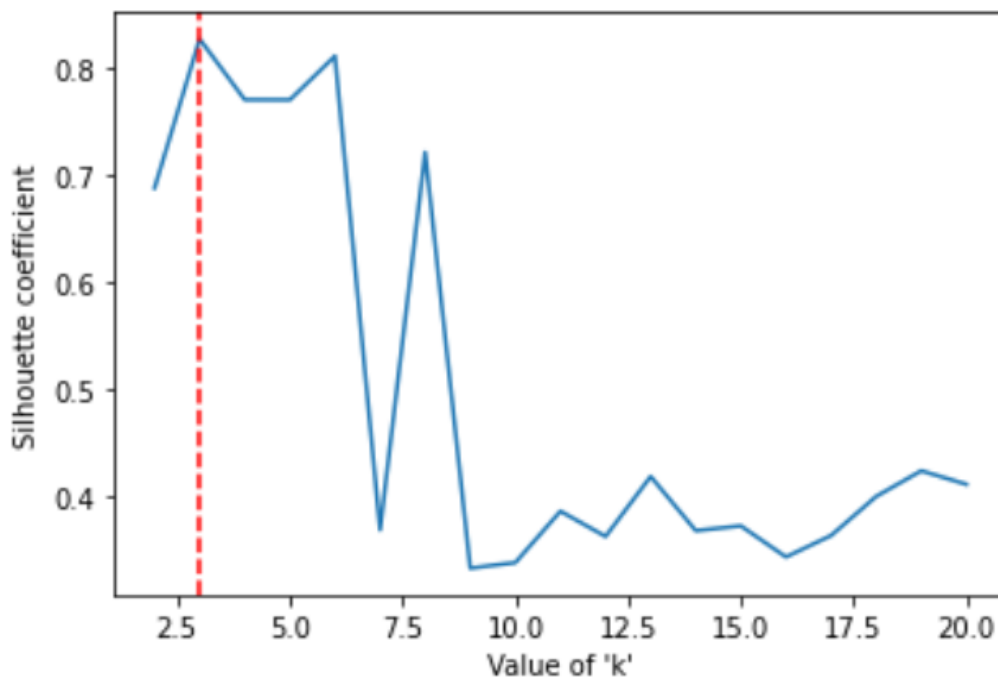
---

## Optimal Number of Clusters

The optimal number of clusters obtained after calculating the silhouette coefficient for  $k = 3, 4, 5$ , and  $6$  is **3**. The values obtained are listed below.

```
Silhouette Coefficient for k = 3 is 0.8700689668748269
Silhouette Coefficient for k = 4 is 0.8629278588557886
Silhouette Coefficient for k = 5 is 0.8658016516543212
Silhouette Coefficient for k = 6 is 0.24924153711362315
Optimal number of clusters = 3
```

The graph below records silhouette coefficients for various values of  $k$ . The dashed line is of  $k = 3$ , which has the maximum value of silhouette coefficient.

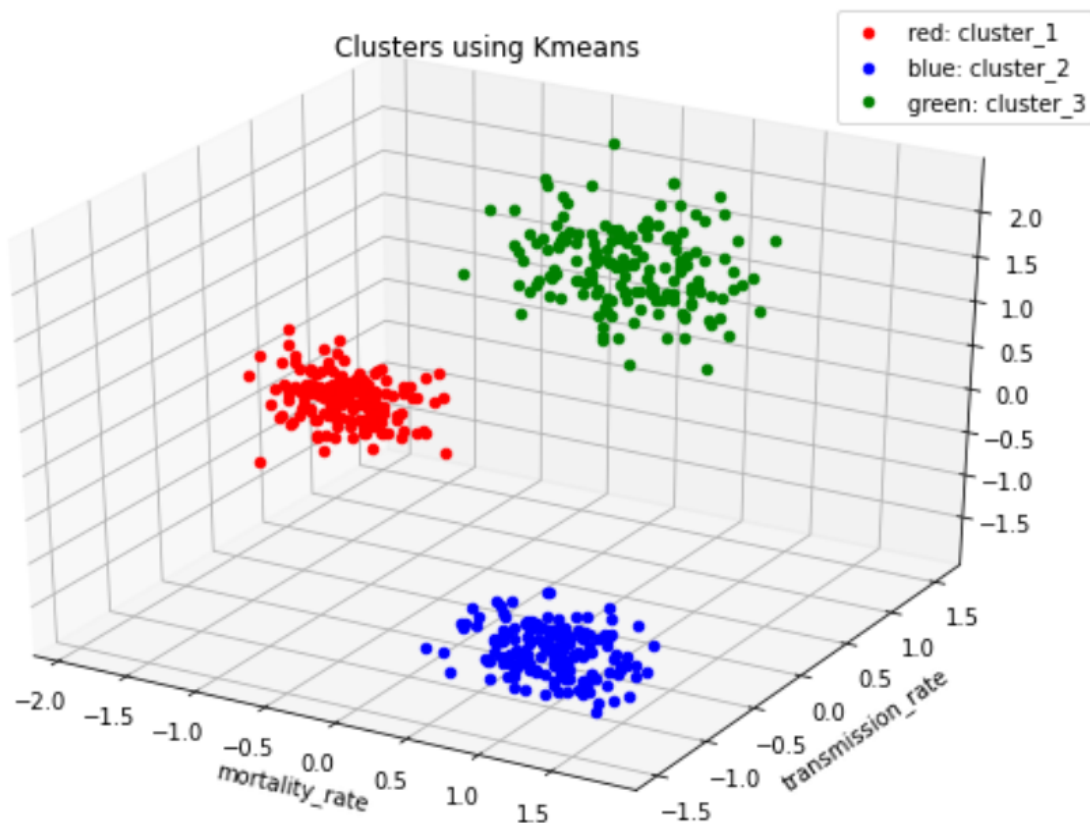


*Graph of Silhouette coefficient vs Value of 'k'*

---

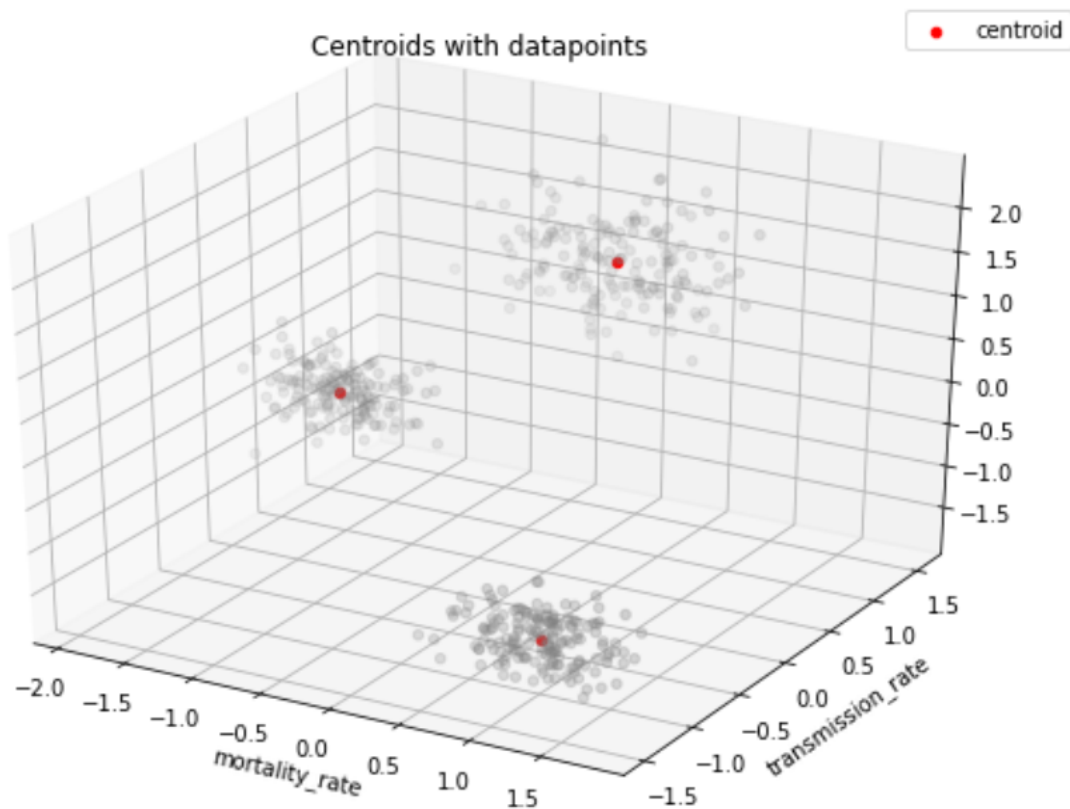
## Analysis of Silhouette Coefficient (in step - 2)

The silhouette coefficient is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. The values closer to one are considered good. In my case, I obtained a silhouette coefficient of `0.8700689668748269` for  $k = 3$  using k-means Clustering. The value is close to one and hence is good. We can also visualize the clusters and centroids for the algorithm.



*Scatter plot of the clusters categorized by K-means*

We can see that my algorithm has categorized all three clusters perfectly. Hence we got a good value of silhouette coefficient. But the value is not exactly 1 because the silhouette coefficient depends upon both **cohesion** and **separation** of clusters and we can see that the clusters are somewhat spread out hence leading to a decrement in cohesion value and therefore the whole coefficient.



*Scatter plot of datapoints with centroids marked*

### **Analysis of Jaccard Similarity Scores (in step - 4)**

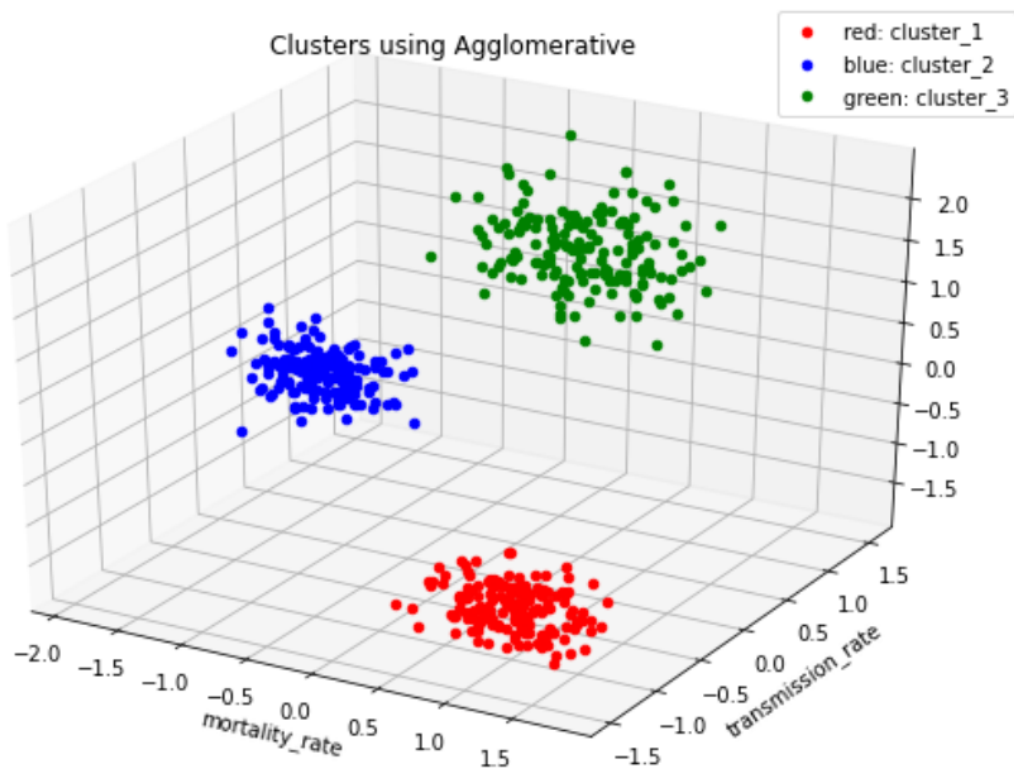
The Jaccard Similarity Scores obtained for all the three mappings from k-means deduced clusters to agglomerative hierarchical clusters is **1**. It is because the Jaccard similarity score is the ratio of intersection of clusters and union of clusters, and the clusters that I received from both algorithms are the same. So  $\text{union}(A,B) = \text{intersection}(A,B)$ , which implies their ratio will be one.

---

Cluster 0 of kmeans is mapped to Cluster 1 of agglomerative  
Cluster 1 of kmeans is mapped to Cluster 0 of agglomerative  
Cluster 2 of kmeans is mapped to Cluster 2 of agglomerative

Jaccard Similarity Score for the 0 -> 1 mapping: 1.0  
Jaccard Similarity Score for the 1 -> 0 mapping: 1.0  
Jaccard Similarity Score for the 2 -> 2 mapping: 1.0

We can also visualize that the clusters obtained from agglomerative are the same as that of k-means.



*Scatter plot of the clusters categorized by Agglomerative clustering*

THE END